AQUATIC ECOSYSTEM RESEARCH

UNIVERSITÄT DUISBURG ESSEN

Offen im Denken

ECOBIAS

2nd Workshop 24.06.2021
University of Tuzla, B&H

Monitoring of aquatic ecosystems using Environmental DNA

Lecturers
Prof. Dr. Florian Leese, Dr. Martina Weiss, Dr. Arne Beermann, M.Sc. Till Macher (University of Duisburg-Essen, Germany

ECOBIAS

Co-funded by the Erasmus+ Programme of the European Union

Environmental DNA datasets

Environmental DNA (eDNA) metabarcoding is a powerful, fast, economically feasible and minimally invasive tool to assess and monitor biodiversity of aquatic ecosystems (lakes, streams, groundwater, sea). The eDNA of a sample can be analyzed for different groups of organisms (such as algae, macroinvertebrates, fish or fungi, often termed “Biological Quality Elements”, BQEs), which is often achieved by using different primers for different groups. Primers are small DNA strands that bind specifically to a region flanking the barcode gene fragment. With two primers, the barcoding fragment can be amplified in a PCR (polymerase chain reaction). The PCR products generated can be sequenced on one of the high-throughput sequencing machines such as Illumina MiSeq, HiSeq, NextSeq or NovaSeq. With the resulting sequence data, the composition of the community can be inferred. Therefore, data first have to be filtered and ‘cleaned’ using specific algorithms (not part of this workshop). After this, sequences are typically clustered in operational taxonomic units (OTUs) that largely represent species diversity. Alternatively, sequence data can be ‘denoised’ so that even genetic variants within species are shown. These sequences are called exact sequence variants (ESVs) or sometimes amplicon sequence variants (ASVs). However, as biologists or environmental managers, we want to know the taxonomic composition of the community and have to make sure that we assign the real taxonomic names to the hundreds and thousands of OTUs or ESVs we obtained.

The focus of this workshop will be on the bioinformatic steps done to assign a species or taxon name to an OTU sequence from an eDNA metabarcoding sequence analysis.

In the following example, eDNA was analyzed to characterize the biodiversity of 4 stream sites. For each sample, 1 L of water was collected, DNA extracted from the sample and subsequently amplified and sequenced. As a result of the bioinformatic processing of the sequences, sequences were clustered based on a 97% similarity threshold into OTUs. An OTU list was compiled, listing the OTUs, the respective OTU sequences and the number of reads per site and number of reads detected in the negative control (Fig. 1. and excel file “OTU list*”). [Please be aware that this is a simplified example and that eDNA datasets normally contain hundreds to thousands of OTUs. See “Outlook” at the end of the document for further information.]*

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	OTU	Sequence	Higher taxonomy	Species	Similarity (Top Hit) [%]	Comment	Sum of reads	OTU occupancy	Site1	Site2	Site3	Site4	NC
2	OTU_1	TATTTTATT							112643	60012	76235	4831	0
3	OTU_2	AATAAATA/							87352	134782	12456	8999	0
4	OTU_3	TTTATTATG							12456	16457	129213	68700	0
5	OTU_4	ACTAAACA/							0	54823	67412	68421	0
6	OTU_5	ATTAAATA/							175210	8301	0	9200	0
7	OTU_6	AATAAATA/							56241	12542	88741	12120	0
8	OTU_7	TGCTTATTC							3612	144101	8211	158	0
9	OTU_8	TACCTATTA							0	17	5112	123123	19
10	OTU_9	TTTAAATAA							0	874	214	9812	0
11	OTU_10	TACCTACTA							0	0	0	4700	0
12	OTU_11	GATGAATA/							0	0	0	3214	0

Fig. 1. OTU list including OTUs, their respective sequences and the number of reads per sampling site (1-4) as well as for a used negative control (NC).

Erasmus + Project No ECOBIAS_609967-EPP-1-2019-1-RS-EPPKA2-CBHE-JP
Development of master curricula in ecological monitoring and aquatic bioassessment for Western Balkans HEIs

Co-funded by the Erasmus+ Programme of the European Union

Taxonomic assignment

1. As a first step, a taxonomic assignment of all OTUs needs to be carried out, using the Barcode of Life Datasystem (BOLD, https://www.boldsystems.org/) database (Fig. 2).

2. Navigate to the "Identification" page, select the database "All Barcode References on BOLD" and copy&paste the sequence of OTU_1 (excel file) into the empty box below (Fig. 3). By pressing "Submit" the first sequence will be compared against the database for taxonomic assignment.

3. Check the result list of your queried sequence against the database (Fig. 4). Every single line in the list represents one reference specimen and sequence deposited in the BOLD database. For every entry, the taxonomy of the reference sequence is presented as well as a "Similarity", which indicates how high the similarity of your queried sequence is compared to the respective reference sequence. Furthermore, in case of published reference sequences, the sequences' metadata can be accessed by clicking on the "blue double-window symbol" right next to "Published".

4.

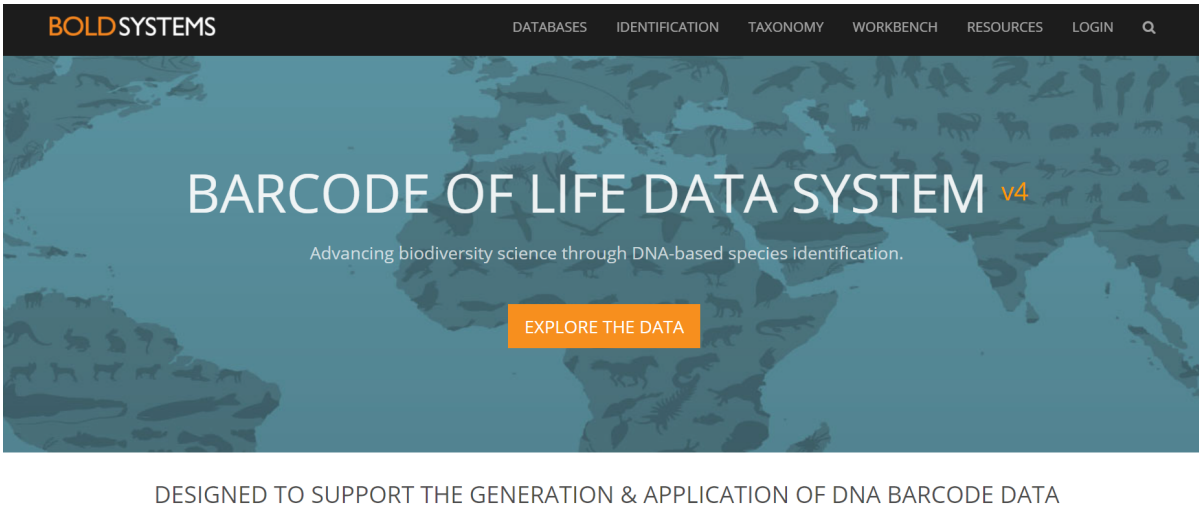


Fig. 2. Interface of the BOLD database (https://www.boldsystems.org/).

Erasmus + Project No ECOBIAS_609967-EPP-1-2019-1-RS-EPPKA2-CBHE-JP
Development of master curricula in ecological monitoring and aquatic bioassessment for Western Balkans HEIs

Co-funded by the Erasmus+ Programme of the European Union

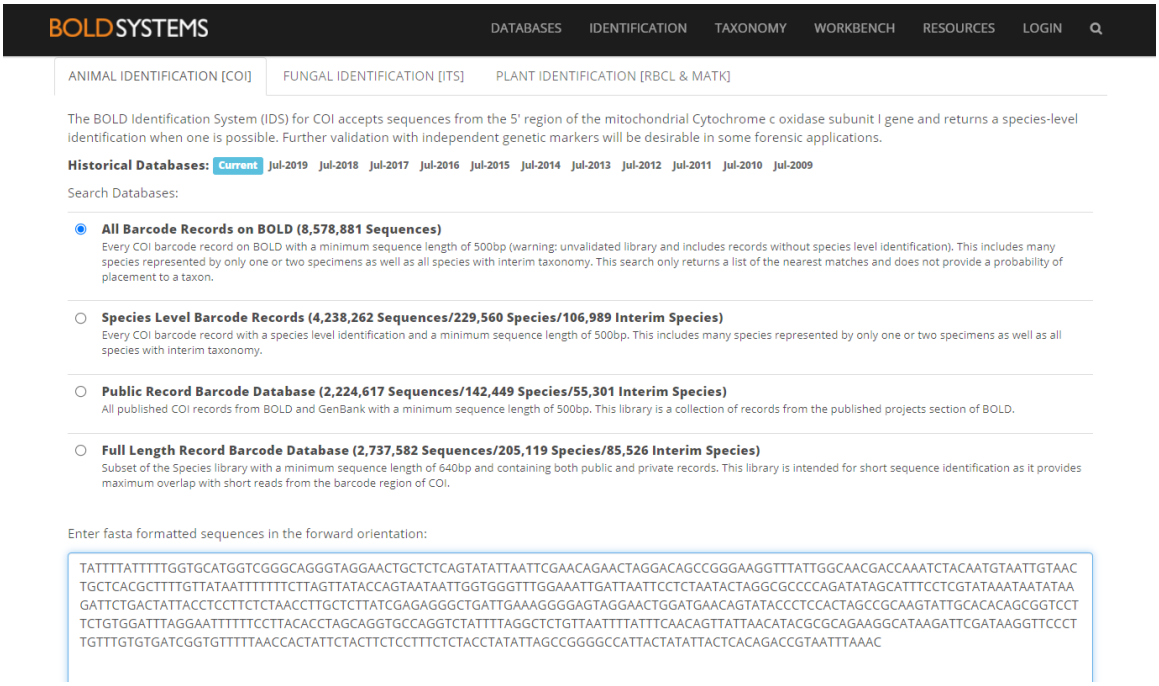
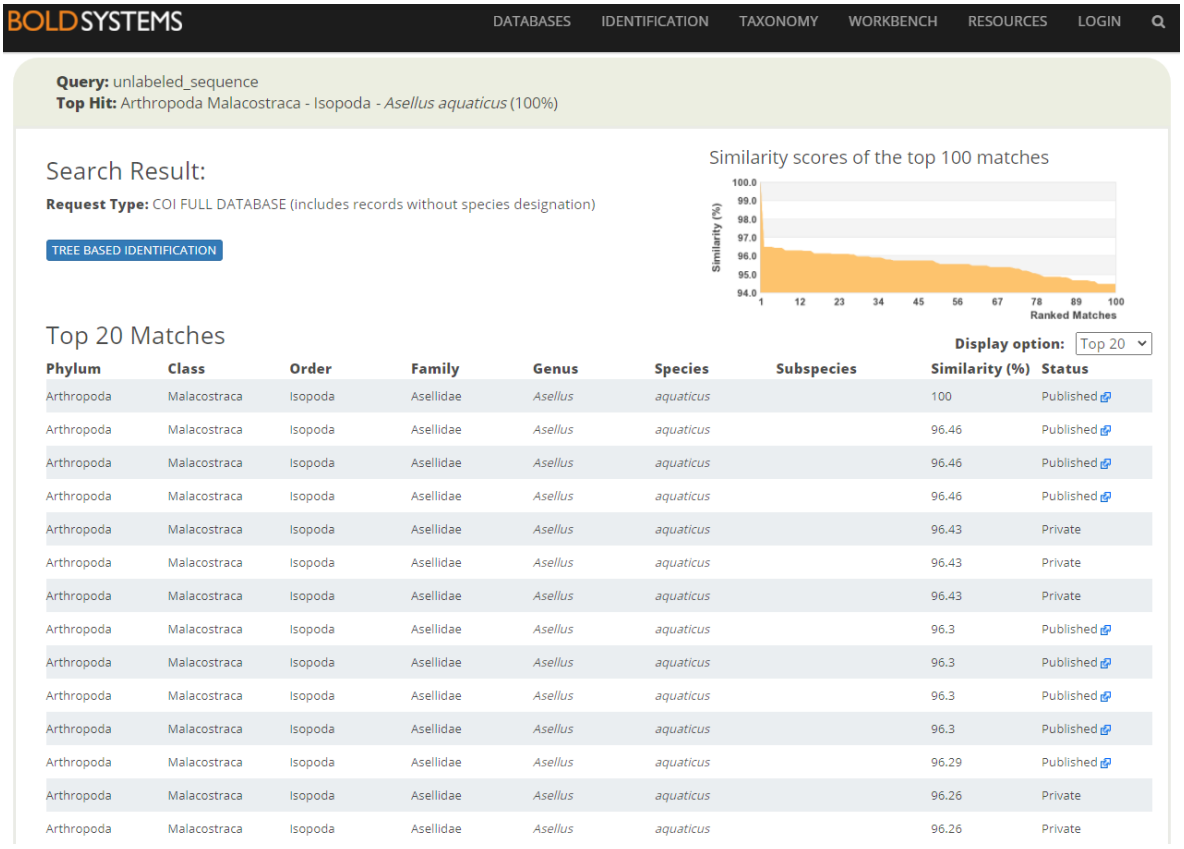


Fig. 3. Interface of the "Identification section". Selecting the "All Barcode Records on BOLD" database will compare a queried sequence against all sequences deposited in BOLD. In contrast, selecting the "Species Level Barcode Records" database will compare a queried sequence against all records with a species level identification.



Phylum	Class	Order	Family	Genus	Species	Subspecies	Similarity (%)	Status
Arthropoda	Malacostraca	Isopoda	Asellidae	*Asellus*	*aquaticus*		100	Published
Arthropoda	Malacostraca	Isopoda	Asellidae	*Asellus*	*aquaticus*		96.46	Published
Arthropoda	Malacostraca	Isopoda	Asellidae	*Asellus*	*aquaticus*		96.46	Published
Arthropoda	Malacostraca	Isopoda	Asellidae	*Asellus*	*aquaticus*		96.46	Published
Arthropoda	Malacostraca	Isopoda	Asellidae	*Asellus*	*aquaticus*		96.43	Private
Arthropoda	Malacostraca	Isopoda	Asellidae	*Asellus*	*aquaticus*		96.43	Private
Arthropoda	Malacostraca	Isopoda	Asellidae	*Asellus*	*aquaticus*		96.43	Private
Arthropoda	Malacostraca	Isopoda	Asellidae	*Asellus*	*aquaticus*		96.3	Published
Arthropoda	Malacostraca	Isopoda	Asellidae	*Asellus*	*aquaticus*		96.3	Published
Arthropoda	Malacostraca	Isopoda	Asellidae	*Asellus*	*aquaticus*		96.3	Published
Arthropoda	Malacostraca	Isopoda	Asellidae	*Asellus*	*aquaticus*		96.3	Published
Arthropoda	Malacostraca	Isopoda	Asellidae	*Asellus*	*aquaticus*		96.29	Published
Arthropoda	Malacostraca	Isopoda	Asellidae	*Asellus*	*aquaticus*		96.26	Private
Arthropoda	Malacostraca	Isopoda	Asellidae	*Asellus*	*aquaticus*		96.26	Private

Fig. 4. Search result of a queried DNA sequence against the BOLD database.

Erasmus + Project No ECOBIAS_609967-EPP-1-2019-1-RS-EPPKA2-CBHE-JP
Development of master curricula in ecological monitoring and aquatic bioassessment for Western Balkans HEIs

Co-funded by the Erasmus+ Programme of the European Union

5. Fill in the “Higher taxonomy”, the “Species” and the “Similarity” of the top hit in the excel file. Add any noteworthy comments on the result of your taxonomic assignment (Fig 5.).
6. Repeat steps 1-4 for all OTUs.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	OTU	Sequence	Higher taxonomy	Species	Similarity (Top Hit) [%]	Comment	Sum of reads	OTU occupancy	Site1	Site2	Site3	Site4	NC
2	OTU_1	TATTTTATT	Crustacea	Asellus aquaticus	100				112643	60012	76235	4831	
3	OTU_2	AATAAATA							87352	134782	12456	8999	
4	OTU_3	TTTATTATG							12456	16457	129213	68700	
5	OTU_4	ACTAAACA							0	54823	67412	68421	
6	OTU_5	ATTAAATA							175210	8301	0	9200	
7	OTU_6	AATAAATA							56241	12542	88741	12120	
8	OTU_7	TGCTTATTC							3612	144101	8211	158	
9	OTU_8	TACCTATTA							0	17	5112	123123	1
10	OTU_9	TTTAAATAA							0	874	214	9812	
11	OTU_10	TACCTACTA							0	0	0	4700	
12	OTU_11	GATGAATA							0	0	0	3214	

Fig. 5. OTU list including taxonomic assignment of OTU_1.

First data analysis steps

As a first insight into your eDNA data subsequent to the taxonomic assignment, overview statistics can be explored.

1. Calculate the sum of reads for every single OTU by using the “=SUM” function in excel (Fig. 6). As a result of most bioinformatic pipelines, the OTU with the highest number of reads is OTU_1. Note that negative controls are usually not included here, but evaluated separately.

G	H	I	J	K	L	M
Sum of reads	OTU occupancy	Site1	Site2	Site3	Site4	NC
=SUM(I2:L2)		112643	60012	76235	4831	0
		87352	134782	12456	8999	0
		12456	16457	129213	68700	0
		0	54823	67412	68421	0
		175210	8301	0	9200	0
		56241	12542	88741	12120	0
		3612	144101	8211	158	0
		0	17	5112	123123	19
		0	874	214	9812	0
		0	0	0	4700	0
		0	0	0	3214	0

Fig. 6. Sum of reads.

2. Calculate the occupancy for every single OTU by using the “=Countif” function in excel (Fig. 7). Analyzing the occupancy provides a general idea, if detected OTUs and species are common or rare.

Erasmus + Project No ECOBIAS_609967-EPP-1-2019-1-RS-EPPKA2-CBHE-JP
Development of master curricula in ecological monitoring and aquatic bioassessment for Western Balkans HEIs

Co-funded by the Erasmus+ Programme of the European Union

H	I	J	K	L	M
OTU occupancy	**Site1**	**Site2**	**Site3**	**Site4**	**NC**
=COUNTIF(I2:L2;">0")	60012	76235	4831		0
	87352	134782	12456	8999	0
	12456	16457	129213	68700	0
	0	54823	67412	68421	0
	175210	8301	0	9200	0
	56241	12542	88741	12120	0
	3612	144101	8211	158	0
	0	17	5112	123123	19
	0	874	214	9812	0
	0	0	0	4700	0
	0	0	0	3214	0

1R x 4C

Fig. 7. OTU occupancy.

3. Calculate the OTU richness per site by using the "=Countif" function in excel (Fig. 8). The OTU richness can be used as a proxy for species richness. As a more conservative approach, only OTUs with a taxonomic assignment on species level can be used to calculate species richness.

H	I	J	K	L	M
OTU occupancy	**Site1**	**Site2**	**Site3**	**Site4**	**NC**
	112643	60012	76235	4831	0
	87352	134782	12456	8999	0
	12456	16457	129213	68700	0
	0	54823	67412	68421	0
	175210	8301	0	9200	0
	56241	12542	88741	12120	0
	3612	144101	8211	158	0
	0	17	5112	123123	19
	0	874	214	9812	0
	0	0	0	4700	0
	0	0	0	3214	0
Richness	=COUNTIF(I2:I12;">0")				

Fig. 8. OTU richness.

Erasmus + Project No ECOBIAS_609967-EPP-1-2019-1-RS-EPPKA2-CBHE-JP
Development of master curricula in ecological monitoring and aquatic bioassessment for Western Balkans HEIs

Co-funded by the Erasmus+ Programme of the European Union

You should now be able to answer following questions:

1. What species were detected in the stream ecosystem sample?
2. Were all OTUs assigned to distinct species with high confidence?
3. Were any species unusual for stream ecosystem datasets? If so, what can be the source of their signal (what do the number of reads in the negative controls indicate)?
4. Do OTUs with higher total read counts always occur more often than OTUs with lower read counts? If not, why is that not the case?
5. Do sample sites 1-4 differ in their OTU richness? Is OTU richness informative even without species-level assignments?

Outlook

As you are already aware, eDNA datasets contain many more OTUs and samples than in the example presented above. As it is often not feasible to manually assign every single OTU to a taxon, a taxonomic assignment step is carried out automatically in many metabarcoding pipelines or separate programs. For instance, the BOLDigger program (https://github.com/DominikBuchner/BOLDigger) automatically queries metabarcoding datasets of hundreds or thousands of OTU sequences against the BOLD database and downloads the "Top 20 Hits" for every OTU sequence including information on the taxonomy (phylum, class, order, family, genus, species and subspecies) as well as the respective similarity values. As a subsequent filtering step, the pipeline assigns different taxonomic levels (species, genus, family, order) to OTUs based on the similarity of reference sequences (98%, 95%, 90%, 85%) and takes the most common name in case of conflicting results (JAMP pipeline criteria, https://github.com/VascoElbrecht/JAMP). Additionally, BOLDigger offers the option to flag suspicious hits, e.g. when reference sequences with high similarity values but two or more different species names are present, or in case the most likely assignment result is represented by only one reference sequence. While this approach of taxonomic assignment is straightforward and fast, it can be inaccurate in certain cases. Keep in mind, that bioinformatic processing, including automatic taxonomic assignment, does not replace carefully verifying and checking assignments and interpreting the data.

While the example presented in this workshop is simplified, the respective analyses steps are the same for larger datasets. The excel file "S2_Invertebrate_OTUs" contains a real metabarcoding dataset from a multiple stressor experiment. Even though this dataset contains 128 samples (C1-C64, L1-L64) and 435 OTUs, you will notice that the same analysis steps can be carried out in this dataset. Feel free to experiment with the data.

Erasmus + Project No ECOBIAS_609967-EPP-1-2019-1-RS-EPPKA2-CBHE-JP
Development of master curricula in ecological monitoring and aquatic bioassessment for Western Balkans HEIs

Co-funded by the Erasmus+ Programme of the European Union

Erasmus + Project No ECOBIAS_609967-EPP-1-2019-1-RS-EPPKA2-CBHE-JP
Development of master curricula in ecological monitoring and
aquatic bioassessment for Western Balkans HEIs

Co-funded by the
Erasmus+ Programme
of the European Union